Application of machine learning methods for cost prediction of drought in France

Antoine Heranval, Olivier Lopez, Maud Thomas

Noname manuscript No.
(will be inserted by the editor)

Application of machine learning methods for cost prediction of drought in France

Antoine HERANVAL[1,2*] · **Olivier LOPEZ**[2] · **Maud THOMAS**[2]

Received: date / Accepted: date

Abstract This paper deals with the prediction of the total amount of a drought episode under the French "Catastrophe Naturelle" regime. Due to the specificity of this regime, a quick prediction of the final amount of an incident is particularly strategic. The approach that we use is based on a database constituted by the French Federation of Insurers in order to cover approximately 70% of the French market. Linking it with meteorological data and socioeconomic data allows to increase our vision of the exposure. Although the database is large, with a wide vision of the French metropolitan territory, data is imbalanced since a large majority of cities are not stroke by catastrophic events. Machine learning methods are used to compute a prediction.

Keywords Natural Catastrophe · Generalized Linear Models · Lasso and Elastic-Net penalties · Extreme Gradient Boosting · Random Forests

Mathematics Subject Classification (2020) MSC code1 · MSC code2 · more

1 Introduction

According to Fédération Française des Assurances (French Federation of Insurance Companies), the cost of the damages caused by natural catastrophes, such as drought, is expected to increase in the following years in France [3]. This is, mainly, due to the general growth of wealth in France and to climate change. Climate change has an important impact on drought and its effects, especially on individual houses due to clay-related subsidence. Although not making the front page, this risk is responsible for about 30% of the total amount of claims paid by the French regime CatNat (Régime d'indemnisation des catastrophes naturelles) [5]. This represents over 11 billions euros with some extreme years, like 2003 where the cost rises up to 2 billion euros. The rise of the number of recent drought events is apparent: out of the six most costly events since the creation of the regime, three have occurred after 2010 [5]. The clay-related subsidence is caused by clay shrinking and swelling, responding to wetting and drying conditions. This leads to vertical and horizontal ground movements, caused by volumetric changes in soil mass, which can

[1]Mission Risques Naturels, 1 rue Jules Lefebvre 75009 Paris, France
[2]Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, 75005 Paris, France
*Corresponding author
E-mails : antoine.heranval@sorbonne-universite.fr, olivier.lopez@sorbonne-universite.fr, maud.thomas@sorbonne-universite.fr

cause significant damage to infrastructure and especially individual houses [2,12]. These damages have been reported in other countries as well and the associated cost is also very high, up to £500 millions per annum in the UK, for instance [25].

In this paper, we propose a methodology to estimate the cost of the consequences of drought shortly after its occurrence for the entire French market. The main goal is to provide tools for insurance companies to assess the severity of future drought events. Due to the large engaged amounts, the potential involvement of the government, or the quality of prevention and assistance delivered to policyholders, evaluating the order of magnitude of the cost of such an episode is a challenging task.

The specificity of the French system, where the "Régime Catastrophe Naturelle" (CatNat regime) applies, makes it even more difficult. Indeed, a very specific administrative procedure—involving insurance companies and the government—governs the compensation and the management of such claims. Financial assistance is then triggered by a decree, which is published only after an examination process that can take a significant amount of time. The approach we develop in this paper aims at predicting if a geographic area (namely a city) will be affected or not by such a claim. The knowledge of such an information is of course crucial to determine the amount of reserves required to face such an event. Once a prediction has been made, a cost can be estimated from the number of exposed houses.

Our methodology to predict the cost of drought events relies on the comparison of different statistical methods such as Generalized Linear Models (combined with Lasso and Elastic-Net penalties [20]) with machine learning algorithms, like Random Forests [13,32] and Extreme Gradient Boosting [17]. The calibration of these methods is performed on a large database provided by Mission Risques Naturels, covering approximately 70% of the French market. An important difficulty stands in the fact that the database we use to calibrate these models is imbalanced. Indeed, catastrophic events such as the one we focus on are relatively rare, and most geographical areas are not affected by any claim at all. To improve the performance and to benefit from all the tested models, we propose an aggregation of the outcomes on which we can base new predictions. The predictions obtained from the different models are thus assessed with Precision and Recall curves, F_1- scores and confusion matrices.

The rest of the paper is organized as follows. In Section 2, we describe the framework of this problem and the variable used to estimate the cost. Section 3 is devoted to the general description of the statistical tools we use. In Section 4, these models are applied to a real database. Section 5 presents the results of the cost prediction. The paper ends with a discussion in Section 6.

2 Description of the problem and associated data

This section presents the context of this work and the problem related to the French CatNat regime. A short presentation of this regime is made in Section 2.1. We then describe our classification problem in Section 2.2, which consists in predicting that a city will have a claim. The database, developed by Missions Risques Naturels (MRN), used to perform this prediction is described in Section 2.3, and the covariates in Section 2.4. We particularly focus, in this last section, on a spatial-temporal meteorological index, the Standardized Soil Wetness Index (SSWI), used to characterize the intensity of the drought. The sensitivity of the soil to clay shrinking and swelling relies on the cartography produced by the Bureau de recherches géologiques et minières (BRGM), a French geological and mining research institute [31]. Finally, we summarize the whole methodology that we propose in Section 2.5.

2.1 Short description of the specificity of French CatNat regime

In France, natural catastrophes are insured through a public-private partnership, called the CatNat regime. This French specificity strongly influences claim management. This natural disaster compensation scheme was created by the Law of July 13th 1982, and is based on a solidarity principle: for every contract, the same additional premium insurance rate, fixed by the government, is used to compensate for the losses of natural disasters. The scope of its application is large: for example, it covers floods, mudslides, earthquakes and landslides. However, it does not include storms, hail and snow which are covered by classical insurance. Without going into the functional details of this regime of compensation, it is important to note that before receiving the compensation, a government decree, recognizing a city in the state of natural catastrophe, must be published in the "Journal Officiel", where all laws and legislative events of the French Republic are published. To receive a compensation, the policyholder must make a request to the CatNat Commission.

This decree comes after a request from the mayor of the city, asking to the government to recognize the event as a natural catastrophe. The decree is motivated by an inter-ministerial commission, which assesses the exceptional situation of the event at city level. For drought events, the evaluation is based on the soil type and moisture. This corresponds to the exposition of clay shrinking and swelling (soil type) and the meteorological intensity of the drought in the city (moisture). The exposition to clay shrinking and swelling is publicly available through a fixed cartography produced by the BRGM shown in Figure 2 [31]. To measure the wetness of the soil, the Commission uses an index developed with the French meteorological institute Météo France. Based on the values of this index for several months and on whether the city has clay areas, the decree will recognize the city in a natural catastrophe state or not [7,8,9]. This process can be time-consuming, the average time between the occurrence of the event of drought and the decision is about 18 months [5], which is a long time for both the policyholder and the insurer. Moreover, if there is no request, or if the Commission refuses the request, there will be no compensation from this scheme. In this case, additional coverage can be provided by the insurance company but for the clay-related subsidence in France, it is very rare. Considering the large amounts at stake and the long delay, insurance companies are trying to anticipate the total cost of an event. This is a motivation for estimating the cost of a drought episode soon after its occurrence, which is our aim in this paper.

2.2 A binary classification problem

To be able to predict the cost of a drought event, a first step would be to identify the cities that might be recognized in a natural disaster state. Unfortunately, this is a difficult task since firstly, there is an uncertainty on whether the city will make a request to the Commission and secondly, the meteorological index used by the Commission as one of the criteria is not available early enough to predict the cost of an event shortly after the occurrence.

To overcome this issue, we propose to rather predict the cities that may have a claim, since thanks to a partnership with the MRN, we have access to a database containing the past claims that have occurred in France. We described in details this database in Sections 2.3 and 2.4.

Mathematically speaking, we are dealing with a binary classification problem. We denote $Y \in \{0,1\}$ the response variable and $X \in \mathbb{R}^p$ the covariates. $Y = Y_{ij}$ is equal to 1 if a drought event has occurred in city i in year j. Our goal is thus to estimate $\mathbb{P}[Y = 1 \mid X]$ The results of this prediction problem are then linked with a cost in Section 2.5. In the next sections, we describe the database and the covariates used to address this problem.

2.3 The SILECC database

Thanks to a collaboration with the MRN, we have access to the database SILECC [4]. This database covers about 70% of the French non-life insurance market, and contains the claims of 12 major French insurance companies from 1987 to 2018 and aggregated them to form this database. Every claim was then standardized and geo-localized. This database is very useful for the insurance market and the FFA: it allows to keep track of the claims related to natural hazards and where they have occurred. It is also of general interest, e.g. to update the cartography of clay shrinking and swelling, as mention in Section 2.4.

While the database covers several natural hazards, we focus in this paper on drought events, that is claims related to clay shrinking and swelling. We used the data from 2003 to 2018 to ensure that we have enough claims for each year, since some companies have not contributed for the whole period. This period provides strong representativeness of drought in France and covers major episodes such as the ones observed in 2003 and in 2011. We used the data from 2003 to 2017 for estimation and 2018 were kept as an illustration of our prediction methods.

In the database, claims represents 6% of the total number of cities present in the database. This corresponds in average to 1 948 cities with a claim per year, out of 34 841 cities in mainland France. This ratio is variable during our time period. As we said, for some years we observe a large number of claims, such as in 2003 with 25 % of the total number of cities with claims, 2017 and 2011 with 12 % and 10 % are also important. Figure 1 shows the proportion of cities with claims by year regarding the total number of cities with claims. Year 2018 is not shown because the claim data are not yet available.

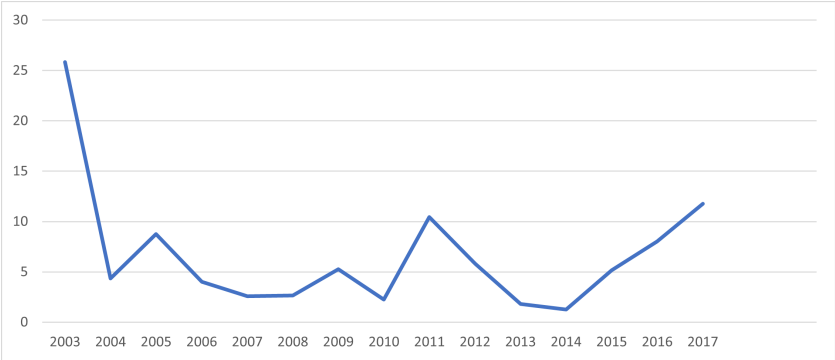

Fig. 1 Proportion of cities with claims by year regarding the total number of cities with claims

2.4 Covariates

To characterize the sensitivity to clay shrinking and swelling in the soil, we use the three classes defined in the cartography done by the BRGM [31]. It is based on the aggregation of three factors, each contributing to the sensitivity of clay-related swelling. The first one is the lithology, which defines the formation and the proportion of clay. The second one is the mineralogical composition of the clay area, especially the proportion of expending minerals (smectites). The third one is the results of the laboratory tests done after the investigation, following a claim for instance. These three factors are rated and then aggregated to supply a susceptibility ranking sensitivity scale. To take the frequency of this phenomenon into account, the sensitivity scale is corrected by the

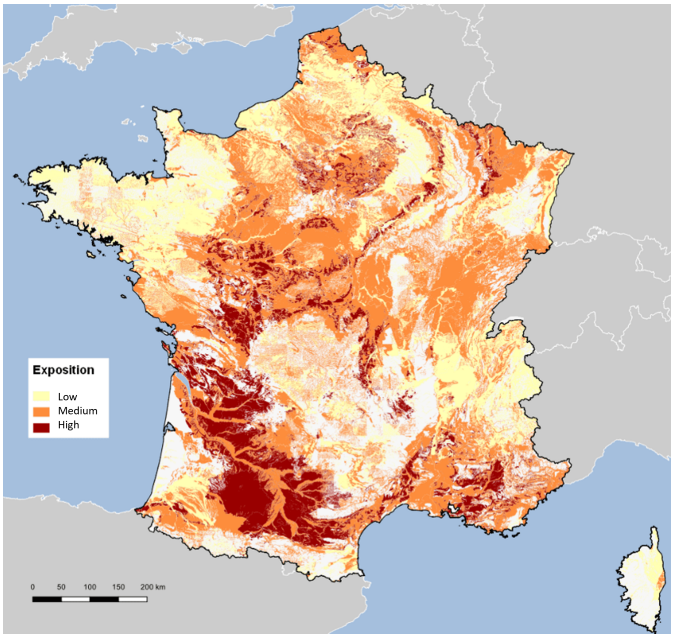

Fig. 2 Cartography of the sensitivity of clay shrinking and swelling clay in France from the BRGM [31] Source: `https://www.georisques.gouv.fr/articles-risques/exposition-du-territoire-au-phenomene`

historical claims. This correction has been updated in 2019, with new claims, with the help of the MRN, as described in [6]. This gives us national cartography at city level, with three values: low, medium and high sensitivity to clay swelling, shown in Figure 2 [31]. This cartography allows us to compute at city level the surface and the proportion of each zone (low, medium and high). We then estimated the number of individual houses in each zone using the data of INSEE in 2015 (French public statistical organization) . To take the evolution of the number of individual houses into account, we applied an augmentation or reduction of 1 % for each year [11].

Regarding the meteorological index, the one used by the Commission has been made available since April 2021 but will not be published early enough for cost prediction. In this respect, we used another spatial-temporal meteorological index, produced by Météo-France, the Standardized Soil Wetness Index (SSWI), as an indicator of the severity of a drought event. This index comes from a research project of Météo-France called Climsec, described in [30], which goal is to assess the impact of climate change on drought in France. The calculation of the SSWI is done through the analysis of precipitation, soil moisture and streamflow outputs from the Safran-Isba-Modcou (SIM) hydrometerological suites [21], and inspired by the Standardized Precipitation Index (SPI) computation procedures [23]. The description of the SSWI is beyond the scope of this paper, a description can be found in [29]. Four time series are then obtained from the SSWI time series as moving averages over one, three, six and twelve months. This gives us four indices for each month representing the wetness of the soil.

The SSWI is a standardized index, its value is thus centered around 0. A negative value suggests a drought whereas a positive one suggests wetness. The value of the index also gives the return period based on the distribution used by Météo-France. In this case, the distribution excludes the extreme values and is calculated on data from 1981 to 2010. Figure 3 shows the geographical distribution of the SSWI for the year 2018. We can see that it is highly variable and that the year 2018 is a year with an important drought in France.

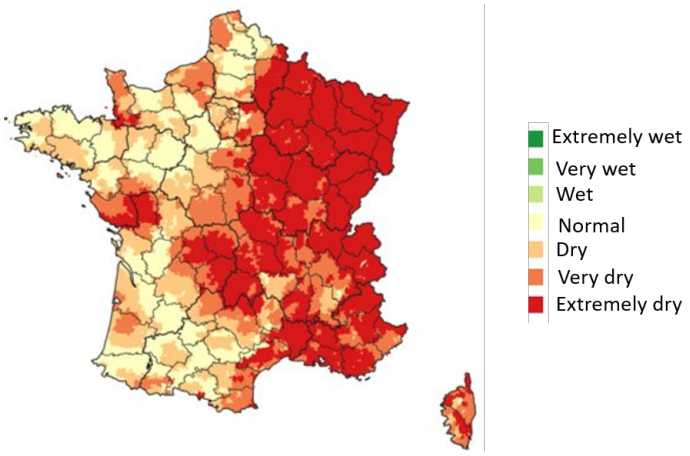

Fig. 3 Cartography of the SSWI for the year 2018. ©MRN. Sources: Mété France SSWI, ADMIN EXPRESS (IGN)

We also computed four other indices, to characterize the drought event itself, as defined in [29]. These indices are computed on the SSWI. The duration of the event corresponds to the number of consecutive months during which the index is negative, its severity to the absolute value of the minimum reached during the event, its magnitude to the absolute value of the sum of the index during the event, and its rarity to a classification of the severity in 7 classes (Extremely wet, Very wet, Wet, Normal, Dry, Very dry and Extremely dry) as shown on Figure 4. These indices are calculated for each city and for each year. In the case of multiple events (in our case the maximum is four events) in one year, we use the value of the indices for all events occurring during the year. Figure 4 illustrates the definition of these indices on an example.

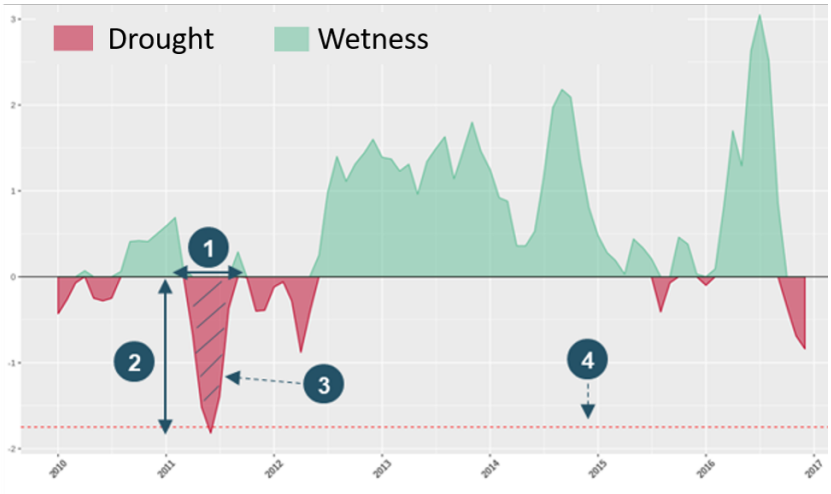

Fig. 4 Description of the used indices based on an example of the SSWI. 1 represents the duration of the event, 2 its severity, 3 its magnitude and 4. its rarity.

We also used indications on the decree of natural catastrophe. Criteria leading to the decree of natural disasters have changed six times over the past 20 years. Therefore the same effects might not lead to the same consequences in our database, depending on the period. All of these constraints make the prediction delicate. We thus implemented a variable that indicates the criteria used by the Commission at the time of the claim.

We also considered the results of the decision of the Commission as if they have used our SSWI meteorological index instead of theirs.

Those variables along with the database described in Section 2.3 constitute our learning database, on which machine learning models, described in Section 3, will be trained.

2.5 Overall methodology

The first step of our method is to predict the cities that will have a claim during a drought event. For that, we used different machine learning models, described in Section 3. Once we know the cities that are likely to be affected by drought, we calculated the number of houses in those cities that are more sensitive to clay shrinking and swelling. To do that we use the cartography done by the BRGM (Figure 2), we count the number of sensitive houses in the city, that is houses localized a zone with medium and strong sensitivity. We then used linear regression to link the number of houses to the cost of the event. This linear model has been trained on our claim database. Figure 5 summarizes this overall methodology.

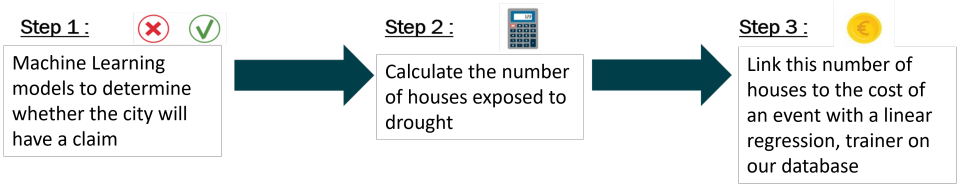

Fig. 5 Overall methodology.

If we write the cost for a city as random variable $Z = Y \times M$, where $Y = 1$ if the city has a claim (and 0 otherwise), and M is the amount of the claim, the classical approach typically assumes that Y and M are independent. This assumption may be questionable. A possibility to avoid it would be, for example to consider a more elaborate model for M as the one we develop below, for example linking M with a large number of covariates (so that Y and M may be independent but conditionally on the values of these covariates). However, this would require more information that we do not have at our disposal.

3 Statistical models

In this section, we describe the different models that we will use and combine to perform the prediction. Section 3.1 is devoted to the Generalized Linear Model with elastic-net estimator, which is a parametric model adapted to covariates with high dimension. Random forests are described in Section 3.2, while a short presentation of boosting methods like XGBoost are done in Section 3.3. We refer to [16], who studied a similar prediction problem with a different database, for additional possible models.

3.1 Generalized linear model with penalty

The Generalized Linear Model (GLM), see for example [24] or [19], is a generic way to consider regression problems which is widely used in insurance. This class of models namely states that, for a response variable Y and $X \in \mathbb{R}^p$ some covariates,

$$g\left(E[Y|X]\right) = X\beta,$$

with $\beta \in \mathbb{R}^p$ is the vector of unknown parameters, and g some monotonous known function, called the link function. Additionally, the conditional distribution of Y given X is assumed to belong to some fixed family of distribution from the exponential family.

In a binary classification problem as ours, the distribution of $Y \mid X$ is assumed to be a Bernoulli distribution with unknown parameter $p(X) = \mathbb{E}[Y|X]$. Regarding the link function g, a standard choice consists in taking $g(y) = \text{logit}(y) = \log\left(y/(1-y)\right)$. This corresponds to the canonical link function (the link function leading to the best theoretical properties in GLM), and also because it is a simple function that maps $[0,1]$ into $\mathbb{R}$.

Estimation can be performed by maximizing the likelihood of the model. However, in our case, the dimension p of the covariates is relatively high. This creates a problem since the statistical precision diminishes if the number of coefficients to estimate is too large. Moreover, many numerical issues can occur. On the other hand, most variables are likely to be irrelevant (but of course, one does not know which by advance). Hence, the GLM elastic-net estimator (GLMNET) is a way to reduce dimension by solving the numerical instability [33].

Let $f_\beta(y,x)$ denote the likelihood of the model. The GLMNET estimator is defined as

$$\widehat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \log(f_\beta(Y_i, X_i)) - \lambda\left\{\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2\right\},$$

with $\|\beta\|_1$ (resp. $\|\beta\|_2$) denotes the l^2-(resp. l^1-) norm of the vector β, the hyper-parameters λ and α being positive constant. The penalization of the log-likelihood by $\|\beta\|_2$ corresponds to a Ridge penalization (see [22]), which stabilizes the estimation result by reducing some numerical issues that may happen in high dimension. On the other hand, the penalization by $\|\beta\|_1$ corresponds to a Lasso penalty (see [28]), designed to produce a sparse model, i.e. a model in which most coefficients of $\widehat{\beta}$ are zero. Hence it allows to reduce the effective dimension of the covariates. The constants λ and α are chosen by cross-validation.

The advantage of GLMNET is to produce an intelligible and easily interpretable model. On the other hand, being able to automatically select the variables having an effect on Y allows to consider a model complex enough to expect a good fit. Nevertheless, the underlying parametric assumption may be too strong in practice, this is why we also turns towards "black-boxes" techniques from machine learning.

3.2 Random Forests

Random Forests (RF) are a machine learning procedure based on the aggregation of regression trees [13]. A regression tree, as introduced by Breiman [14], estimates the function $p(x)$ by

$$\widehat{p}(x) = \sum_{i=j}^{K} p_j R_j(x), \tag{1}$$

where, for all x, $R_j(x) = 0$ for all j except for one. Namely, R_j are "rules" that are associated with a partition of the space of the covariates, allowing to affect x to the unique set of the

partition to which it belongs. In regression trees, these partitions are hyper-rectangles, that is $R_j(x) = 1$ if and only if $x \in \cap_{k=1}^{d}\{x : x_{k,l} \leq x_l \leq x_{k,r}\}$, where $x = (x_1, ..., x_d)$. This partition is obtained iteratively through the CART algorithm, see [14]. The estimation of the values p_j is then done for each region of the space $R_j(x)$.

Regression trees have many appealing properties, like allowing to introduce non-linearities while still producing a model which can be easily understood. However, their main drawback is their instability: new incoming data may disrupt the structure of the partition. RF are a way to stabilize this technique, while allowing to capture more elaborate shapes of regression function $p(x)$. They are obtained by averaging regression trees with some specificities:

- each tree is of small size (A small value of K in Equation (1) is imposed);
- each tree grows on a separate bootstrap sample;
- the rules of a given tree are based only on a small subset of the covariates (subset which is selected randomly).

In the following, we use the implementation of RF from the `R`-package `ranger` [32].

3.3 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBOOST), see [17], is an alternative method to RF which also relies on regression trees, but instead of fitting these trees in a parallelized way, it is done iteratively. The predictor $\widehat{p}^{(t)}(x)$ at the $t-$th step of the algorithm is $\widehat{p}^{(t-1)}(x) + \pi_t(x)$, where $\pi_t(x)$ is a regression tree which is selected in order to make the loss function decrease as much as possible, that is to maximize the log-likelihood in the Bernoulli case with a regularization penalty.

Let $\ell(y_i, \widehat{p}^{(t-1)}(x_i))$ denote the opposite of the log-likelihood for observation i (y_i, x_i) at step $t-1$ (this function is also called cross-entropy in the learning literature). At step t, the algorithm tries to find π_t that maximizes

$$\sum_{i=1}^{n} \partial_2 \ell(y_i, \widehat{p}^{(t-1)}(x_i)) \times \pi_t(x_i) + \frac{1}{2}\partial_2^2 \ell(y_i, \widehat{p}^{(t-1)}(x_i)) \times \pi_t^2(x_i) + pen(\pi_t),$$

where *pen* denotes the regularization penalty, and ∂_2 (resp. ∂_2^2) denotes the (resp. second order) partial derivative of a function with respect to its second argument.

4 Prediction results for the SILECC database

This section presents the prediction results obtained on the database SILECC for the different models described in Section 3. The evaluation of the performance is made through the F_1-score, Precision and Recall curves and confusion matrices, described in Section 4.1. The results are then shown in Section 4.2 we will disclose the results of the different models used.

4.1 Evaluation of the performance

To assess the performance of the different models, we have randomly split our database into a train set (80%) and test set (20%). Recall that our database is very imbalanced in the sense that the proportion of cities that have had a claim is very small.

Common methods to assess the performance of binary classifiers include true positive and true negative rates, and ROC (Receiver Operating Characteristics) curves, which display the true positive rate against the false positive rate. These methods, however, are uninformative when the classes are severely imbalanced. In this context, F_1-score and Precision-Recall curves (PRC) have been shown to be more informative [15,27]. They are both based on the values of

$$\text{Precision}(p_c) = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

against the values of

$$\text{Recall}(p_c) = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}.$$

where p_c is a cut-off probability p_c varying between 0 and 1. Precision quantifies the number of correct positive predictions out all positive predictions made; and Recall (often also called Sensitivity) quantifies the number of correct positive predictions out of all positive predictions that could have been made. Both focus on the Positives class (the minority class, cities with a claim) and are unconcerned with the Negatives (the majority class, cities without a claim).

The F_1-score allows us to combine these two measures in a single index by taking the harmonic mean of those two values. It is derived from the F-Measure introduced in [18] and [26] and defined as

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2}$$

To closer the F_1-score is to 1, the better the prediction model is.

PRC display the values of Precision and Recall as the cut-off probability p_c varies from 0 to 1. The PRC of a skillful model bows towards the point with coordinates $(1, 1)$. The curve of a no-skill classifier will be a horizontal line on the plot with a y-coordinate proportional to the number of Positives in the dataset. For a balanced dataset this will be 0.5 [15].

PRC and F_1-score are complementary in our approach, the PRC is used on the probability outcomes of the models, it gives the best configuration and the best model. Whereas the F_1-score is used to select the best threshold value used in the prediction of the two classes, for each model. We will detail this process in the next section.

4.2 Results

In this section, we present the main results of our analysis for the models of Section 3, we also consider the aggregation of these 3 models. For the aggregation (AGGREGATE), we examined the mean of the three probability outcome of each model, to make a synthesis of all of the predictions. For each, we show ROC and PR curves in Figure 6 and computed the Area Under the Curve (AUC), shown in Table 1. The closer the AUC is to 1 the better is the prediction method .

MODEL	AUC ROC	AUC PRC
GLMNET	0.907	0.503
RF	0.933	0.604
XGBOOST	0.936	0.609
AGGREGATE	0.936	0.615

Table 1 AUC of the different models

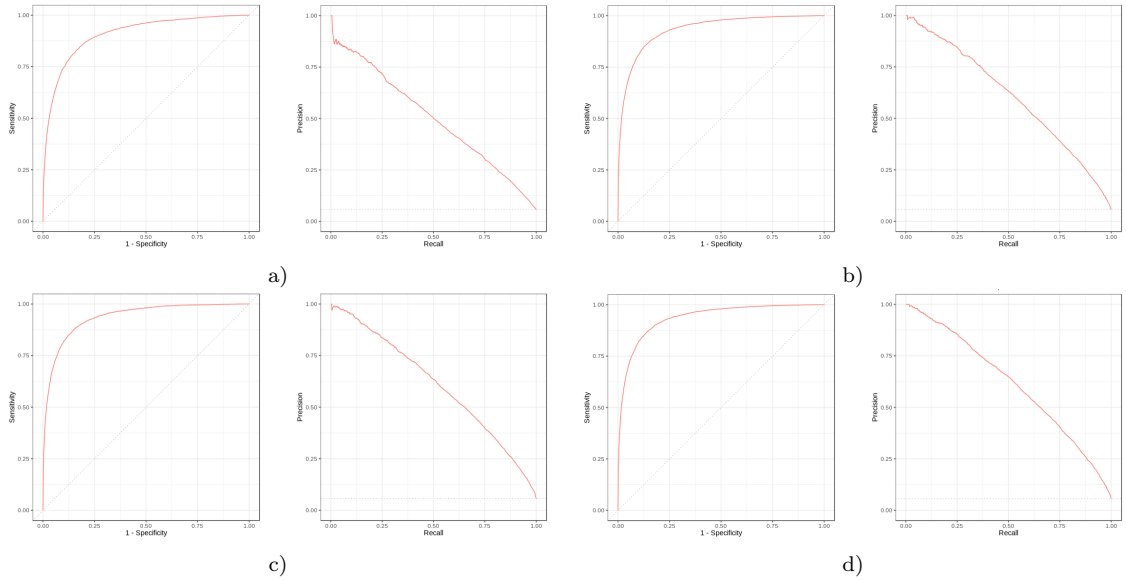

Fig. 6 ROC and PR curves for a) GLMNET b) RF c) XGBOOST d) AGGREGATE on the test sample. This test samples contains 5 924 cities with a claim and 98 596 without a claim.

These results show that the XGBOOST and RF seem to provide better results on our data than the GLMNET. However, the aggregation gives the best results which is a strong advocate for the use of this method. We also see why we used the AUC PRC, if we consider only the AUC ROC we have excellent discrimination because the model successfully predicts a lot of 0. In this case, a model that will predict only 0, will also have a good ROC curve. However, once we use an index that gives more importance to correctly predicting the 1, the results are much more nuanced. We obtained an AUC PRC of around 0.60 which reflects the nature of our problem and the difficulty to predict such claims. Again, we point that ROC curves are not an appropriate tool to assess the performance of our binary classifier since our data are imbalanced.

The PRC seem to be reasonable given our classification problem, they are not perfect but the trade-off is acceptable. The XGBOOST and RF seem to work better, as the beginning of the graph decreases slower in comparison to the one in the GLMNET. It is also very interesting to see that the aggregation smooths the results and takes the best out of each model. The beginning of the graph is less discontinuous. The aggregation appears to be the best model to use for the prediction.

We then selected a threshold to make the prediction, that is the value of the cut-off probability p_c over which we consider the prediction to be a 1. A classical value of 0.5 seems arbitrary as the predictions we try to make are rare and so a probability of 0.5 can already be a strong score. To do that we used the F_1-score and tried different thresholds with a step of 0.001. The results are presented in table 2.

Again we can see that the best F_1-score is obtained by the aggregation of the model, with a value of 0.576 and a threshold of 0.264. We find the same order as before, the XGBOOST works best and the Random Forest. All the thresholds are around 0.25, this confirms the idea that a threshold of 0.5 would have been arbitrary and would not make the best prediction.

Figure 7 shows the confusion matrices for each model. We can note that each model seems to predict the same number of cities with claims, around 6 000 on the test set. This number

MODEL	F1-score	Thresholds
GLMNET	0.503	0.221
RF	0.570	0.306
XGBOOST	0.573	0.291
AGGREGATE	0.576	0.264

Table 2 Best F_1-score and thresholds associated for each model.

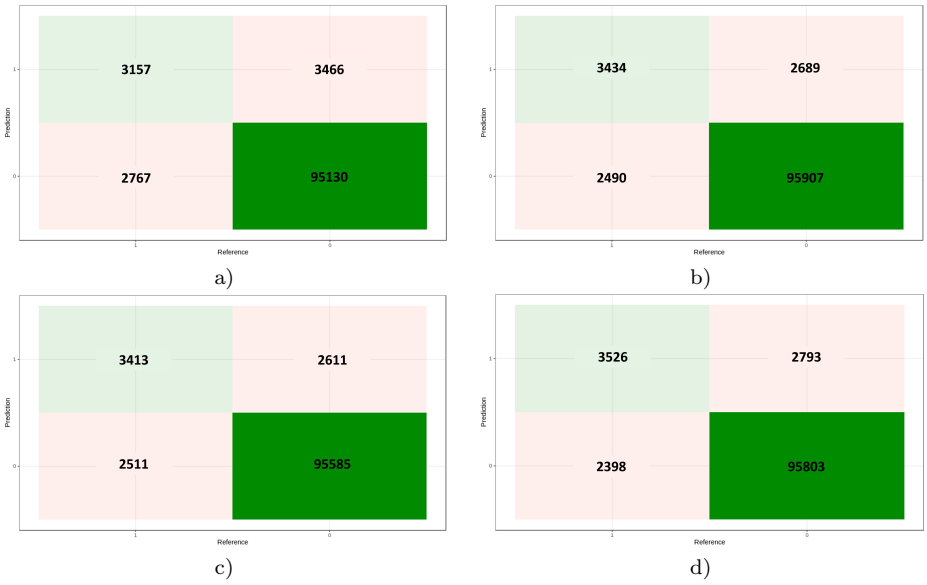

Fig. 7 Confusion matrices for a) GLMNET b) RF c) XGBOOST d) AGGREGATE on the test sample.

is close to the actual number, but there are errors in the predictions. GLMNET does more false predictions, which can explain the differences seen in the F_1-score. When we examine the confusion matrices there are no notable differences between the other models.

5 Cost prediction

Now that we have obtained a model to predict whether a city may have a claim, we try to determine the related cost of such an event. Model for evaluating the cost is described in Section 5.1. The final results are shown in Section 5.2

5.1 Linear regression

As we mentioned above, to estimate the cost of drought, we linked the number of houses to the total cost. The variable used for the cost is the aggregation of all claims for a year, reported in the database SILECC. Since the database represents 70% of the market, we multiplied each cost by 1.42 to obtain a cost for the entire French market. We took the same sample as before, using the data from 2003 to 2017. The maximum cost is in 2003 with 2 billion euros, with an average of 415 million euros per year. We tried to make the link between this global loss, and the number of houses in the cities that are stroke by drought during the corresponding period. The

number of houses has a similar distribution, going up to 4,7 millions in 2003, with an average of 1.7 million houses per year.

We fitted a linear regression on the data and the equation linking those two variables is the following, defining L as the total loss over one year:

$$L = \text{Number of houses} \times 464.4 - 4.121e + 08, \tag{3}$$

with standard errors of $1.15e+08$ for the intercept and 55.79 for the number of houses. We found a good correlation between these two variables in our database, with a $R^2 = 0.84$ and a residual standard error of $2.198e + 08$. Figure 8 shows that the linear regression is a good approximation for our problem. We use the 95% confidence interval, which we will use for our prediction.

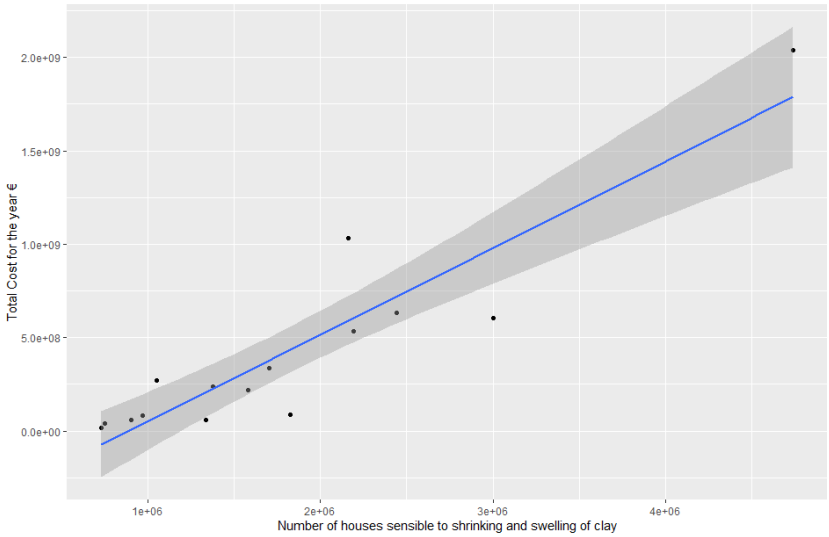

Fig. 8 Linear regression for the number of houses. The points are the observations, the blue line the regression line and the grey are the confidence band.

This linear model is of course very rough, but has to be fitted on the small number of observations that we have (only 15), which explains the choice for the most simple regression model. Although the R^2 is relatively close to 1, one should of course not be overconfident on this fit due to the small number of points used to estimate the parameters of the model.

5.2 Results of the cost prediction for 2018

The previous model is then linked with the prediction models of Section 3. Once we have the number of houses we can estimate a cost with a confidence interval. In our case, the total loss is

$$L = \sum_{i=1}^{N} Y_i M_i,$$

where $Y_i = 1$ if a claim occurred and 0 otherwise, and M_i is the corresponding amount of the claim (for which we know the number of houses $\mathfrak{n}_i$), and N is the number of considered cities. If

Y_i and M_i are independent, then the variance is

$$\mathrm{Var}(Y_i M_i \mid X) = \left(\mathbb{E}[M^2]p_i(1-p_i) + p_i \mathrm{Var}(M)\right),$$

where $p_i = \mathbb{P}(Y_i = 1|X)$. Hence the variance σ^2 of L can be estimated by

$$\widehat{\sigma}^2 = \sum_{i=1}^{n} \left(\hat{m}_{2,i} p_i(1-p_i) + p_i \tilde{\sigma}^2\right),$$

where $\tilde{\sigma}$ is the estimated standard error in the linear regression model of Section 5.1, and

$$\hat{m}_{2,i} = \tilde{\sigma}^2 + (\hat{\alpha} + \hat{\beta}\mathfrak{n}_i)^2,$$

with $(\hat{\alpha}, \hat{\beta}) = (-4.121e+08, 464.4)$, as estimated in the previous section.

Then the 95% confidence intervals of our estimation can be approximated by

$$\widehat{L} \pm 1.96\widehat{\sigma}^2$$

The results of this estimation are displayed in Table 3. The FFA estimated the cost of the drought in France for 2018 to 900 million euros [10], with competing estimations between 1 100 and 1 300 million euros [1]. The outcome of the aggregation model shows the same results. Even if we do not have a very good precision at the city level, the general cost is consistent with the observed data.

MODEL	Lower bound	Estimate	Upper bound
GLMNET	519 031 971	579 350 811	639 669 651
RF	1 505 065 989	1 618 225 685	1 731 385 382
XGBOOST	906 768 050	977 086 655	1 047 405 261
AGGREGATE	879 561 915	965 750 651	1 051 939 387

Table 3 Estimates and confidence intervals of the predicted costs for 2018 (in euros)

The difference in the cost predicted by the models is essentially due to the predicted number of houses. This is also due to the number of cities predicted by each model, as we can notice in Table 4. We can note for 2018 that RF predict more cities with a claim whereas GLMNET predicts less. This once again advocates for the use of the aggregation because it has an averaging effect on the prediction and allows us to take the best of each prediction.

MODEL	Number of cities	Number of sensitive houses
GLMNET	1 364	2 134 000
RF	5 525	4 371 000
XGBOOST	1 800	2 991 000
AGGREGATE	1 823	2 966 000

Table 4 Predicted number of sensitive houses and cities for the 2018-drought

6 Conclusion and discussions

In this work, we developed a method to predict the cost of the consequences of drought for the entire French market. We first used a Generalized Linear Model with Elastic-Net penalization, Random Forest and Extreme Gradient Boosting models with different discriminant thresholds to predict the cities that may have a claim. Based on these predictions, we calculated the number of houses sensitive to the clay shrinking and swelling and then computed the total cost through a linear regression.

We obtained encouraging results for such a complex phenomenon, although lots of uncertainties remain. A database of all claims due to natural catastrophe can be used to estimate the cost of an event. Despite moderate results for the prediction of the impacted cities, we have coherent results when it comes to the cost prediction. The used database, the process of natural catastrophe and the nature of this hazard make the modeling very complex and uncertain. Indeed, our database is based on past claims, reported by different insurers and can contain imprecision, that can affect the results of the impacted city.

The second difficulty is the process of decrees of natural disasters. To be able to obtain compensation, and therefore to appear in our database, the city must be recognized by the Commission, there may be claims in cities that were not recognized. Our models can find such claims, but we are not able to tell whether it is true or not. Also, there have been in the past 20 years, six changes in the criteria, those changes have induced changes in the meteorological characteristics of the cities affected by drought. Therefore, in our train database, we can have different characteristics that will have different effects depending on the criterion.

Moreover, with the meteorological and geological variables that were at our disposal, we only addressed a part of the factors that drive the risk. The interaction between the structure of the house and the composition of the soil plays an important role to determine whether the house will be damaged by a drought event. We took the nature of the soil into account with the BRGM indicators but it is very difficult to take the structure of the house into account due to the lack of data on the different types of foundations, especially at a local level.

We also faced difficulties to assess our model. As mention above there is uncertainty on the results due to the recognition process. More generally it is difficult to find the right score to judge a model, especially with imbalanced data. Furthermore, the prediction that we make can only be verified for one or two years, and even more if we want to have all the claims. The results are encouraging but need to be consolidated by more accurate predictions.

Despite these difficulties, the methods we developed allowed us to improve the prediction of the losses from drought in this particular French context. The techniques we used could be improved with additional amount of data, and with additional knowledge on the spatial dependence phenomena between cities (namely how two close cities may coordinate or not their responses). Let us point that the main advantage of our approach is to provide a fast answer to the question of the cost of such natural event, in a context where the speed of reaction is important to optimize risk management. Finally, let us mention that the methods we developed could also be extended to approximate or predict the index used by the CatNat Commission, in order to improve the prediction. Indeed, this index has recently been made available.

Declarations

Funding This research was supported by the Mission Risques Naturels.

Conflict of interest The authors declare that they have no conflict of interest.

Availability of data and material The database used is not publicly available.

Code availability The code is not publicly available.

References

1. Catastrophes naturelles : la facture salée des sécheresses à répétition. Tech. rep., Argus de l'Assurance. URL https://www.argusdelassurance.com/assurance-dommages/catastrophes-naturelles-la-facture-salee-des-secheresses-a-repetition.169969
2. Avant de construire – prendre en compte les risques du terrain. Tech. rep., Agence Qualité Construction (2014). URL https://qualiteconstruction.com/publication/avant-de-construire-prendre-en-compte-les-risques-du-terrain/
3. Etude : Changement climatique et assurance à l'horizon 2040. Tech. rep., Fédération Française de l'assurance (2015). URL https://www.ffa-assurance.fr/la-federation/publications/enjeux-climatiques/etude-changement-climatique-et-assurance-horizon-2040
4. Présentation de la MRN. Tech. rep., Mission Risques Naturels (2018). URL https://www.mrn.asso.fr/wp-content/uploads/2018/09/presentation-mrn_v21092018-1.pdf
5. Sécheresse géotechnique, de la connaissance de l'aléa à l'analyse de l'endommagement du bâti. Tech. rep., Mission Risques Naturels (2018). URL https://www.mrn.asso.fr/wp-content/uploads/2019/01/21-01-2018_rapport-mrn_secheresse-2018.pdf
6. Lettre d'information de la Mission Risques Naturels 30, juillet 2019. Tech. rep., Mission Risques Naturels (2019). URL https://www.mrn.asso.fr/wp-content/uploads/2019/10/lettre-n30_vf.pdf
7. Procédure de reconnaissance de l'état de catastrophe naturelle - Révision des critères permettant de caractériser l'intensité des épisodes de sécheresse-réhydratation des sols à l'origine de mouvements de terrain différentiels. Tech. rep., Ministère de l'interieur (2019). URL https://www.legifrance.gouv.fr/download/pdf/circ?id=44648
8. Contribution de Météo-France à l'analyse de la sécheresse géotechnique à l'attention de la Commission CatNat pour l'année 2019. Tech. rep., Météo France, Direction de la Climatologie et des Services Climatiques (2020). URL http://www.meteofrance.fr/documents/10192/36885873/Rapport-CatNat-Secheresse-2020.pdf
9. Météo-France dans le dispositif CATNATsécheresse. Tech. rep., Météo France (2020). URL http://www.meteofrance.fr/documents/10192/79826318/Meteo-France+dans+le+dispositif+CATNAT+secheresse
10. L'assurance des événements naturels en 2019. Tech. rep., Fédération Française de l'assurance (2021). URL https://www.mrn.asso.fr/wp-content/uploads/2021/03/2021-mrn-lassurance-des-evenements-naturels-en-2019.pdf
11. Arnold, C.: Le parc de logements en france au 1er janvier 2018. Tech. rep., INSEE (2018). URL https://www.insee.fr/fr/statistiques/3620894
12. Assadollahi, H.: The impact of climatic events and drought on the shrinkage and swelling phenomenon of clayey soils interacting with constructions. Ph.D. thesis, Université de Strasbourg (2019). URL https://tel.archives-ouvertes.fr/tel-02331567/file/Assadollahi_Hossein_2019_ED269.pdfs
13. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001). DOI 10.1023/A:1010933404324
14. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. CRC press (1984). DOI 10.1201/9781315139470
15. Brownlee, J.: Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. Machine Learning Mastery (2020)
16. Charpentier, A., James, M., Ali, H.: Predicting drought and subsidence risks in france URL http://arxiv.org/abs/2107.07668
17. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM. DOI 10.1145/2939672.2939785
18. Chinchor, N., Sundheim, B.M.: Muc-5 evaluation metrics. In: Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993 (1993). URL https://aclanthology.org/M93-1007.pdf
19. Denuit, M., Charpentier, A.: Mathematiques de l'Assurance Non-Vie. Tome II: Tarification et Provisionnement (2005)
20. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of statistical software **33**(1), 1 (2010). URL https://pubmed.ncbi.nlm.nih.gov/20808728/
21. Habets, F., Boone, A., Champeaux, J.L., Etchevers, P., Franchistéguy, L., Leblois, E., Ledoux, E., Le Moigne, P., Martin, E., Morel, S., Noilhan, J., Quintana Seguí, P., Rousset-Regimbeau, F., Viennot, P.: The SAFRAN-ISBA-MODCOU hydrometeorological model applied over france **113**, D06113. DOI 10.1029/2007JD008548

22. Marquardt, D.W., Snee, R.D.: Ridge regression in practice. The American Statistician **29**(1), 3–20 (1975). DOI 10.1080/00031305.1975.10479105
23. McKee, T.B., Doesken, N.J., Kleist, J., et al.: The relationship of drought frequency and duration to time scales. In: Proceedings of the 8th Conference on Applied Climatology, vol. 17, pp. 179–183. Boston (1993). URL `https://climate.colostate.edu/pdfs/relationshipofdroughtfrequency.pdf`
24. Nelder, J.A., Wedderburn, R.W.: Generalized linear models. Journal of the Royal Statistical Society: Series A (General) **135**(3), 370–384 (1972). DOI 10.2307/2344614
25. Pritchard, O.G., Hallett, S.H., Farewell, T.S.: Probabilistic soil moisture projections to assess Great Britain's future clay-related subsidence hazard **133**(4), 635–650. DOI 10.1007/s10584-015-1486-z
26. Rijsbergen, C.: Information retrieval 2nd ed buttersworth. London (1979)
27. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets **10**(3), e0118432. DOI 10.1371/journal.pone.0118432
28. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996). DOI 10.1111/j.2517-6161.1996.tb02080.x
29. Vidal, J.P., Martin, E., Kitova, N., Najac, J., Soubeyroux, J.M.: Evolution of spatio-temporal drought characteristics: validation, projections and effect of adaptation scenarios **16**(8), 2935–2955. DOI 10.5194/hess-16-2935-2012
30. Vidal, J.P., Moisselin, J.M.: Impact du changement climatique sur les sécheresses en France (2011). URL `http://www.drias-climat.fr/public/shared/rapport_final_CLIMSEC.pdf`
31. Vincent, M., Plat, E., Le Roy, S.: Cartographie de l'aléa retrait-gonflement et plans de prévention des risques (120), 189–200. DOI 10.1051/geotech/2007120189
32. Wright, M.N., Ziegler, A.: ranger: A fast implementation of random forests for high dimensional data in c++ and r **77**(1). DOI 10.18637/jss.v077.i01
33. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) **67**(2), 301–320 (2005). DOI 10.1111/j.1467-9868.2005.00503.x